雑音・残響下におけるRahmonicとメルケプストラムを用いた叫び声検出

AUTHOR(S):

福森, 隆寛; 中山, 雅人; 西浦, 敬信; 南條, 浩輝

CITATION:

福森, 隆寛 ...[et al]. 雑音・残響下におけるRahmonicとメルケプストラムを用いた叫び声検出. 電子情報通信学会技術研究報告 2017, 117(189): 49-54: SP2017-31.

ISSUE DATE:

2017-08

URL:

http://hdl.handle.net/2433/229410

一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2017-31(2017-08)

雑音・残響下における Rahmonic とメルケプストラムを用いた叫び声検出

福森　隆寛† 　中山　雅人† 　西浦　敬信† 　南條　浩輝††

† 立命館大学 情報理工学部 〒 525–8577 滋賀県草津市野路東 1-1-1
†† 京都大学 学術情報メディアセンター 〒 606–8501 京都府京都市左京区吉田二本松町
E-mail: †{fukumori@fc, mnaka@fc, nishiura@is}.ritsumei.ac.jp, ††nanjo@media.kyoto-u.ac.jp

あらまし　本稿では，雑音・残響環境下における Rahmonic とメルケプストラム（Mel-Frequency Cepstrum Coefficients: MFCCs）を用いた叫び声検出手法について述べる．人間の聴覚特性を考慮したケプストラム係数である MFCCs は，音韻を特定するための声道特徴量を示しており，また基本周波数の低調波成分である Rahmonic は，人間の声帯運動に関わる特徴を表現している．これまで，我々は大量の平静音声と叫び声から抽出した MFCCs と Rahmonic に基づいて構築した 3 種類の音響モデル（GMM: Gaussian Mixture Model，HMM: Hidden Markov Model，DNN: Deep Neural Network）を用いて叫び声検出手法の有効性を示していた．特に前報までは，クリーン環境と雑音環境における叫び声の検出性能を評価し，提案手法の高い検出性能を確認した．本稿では，更に実環境を想定して，雑音と残響が混在する環境において叫ばれた音声の検出性能を評価する．評価実験の結果，MFCCs と Rahmonic を音声特徴量として用いることで，雑音や残響の種類や SNR によらず，叫び声の発声機構（声道特性と声帯特性）を効率よく表現できることを確認した．また，ほとんどの騒音・雑音環境において音響モデルとして DNN を用いることで GMM や HMM よりも高い叫び声検出性能を達成できた．

キーワード　叫び声検出，雑音・残響環境，Rahmonic，メルケプストラム

Detection of noisy-and-reverberant shouted speech using rahmonic and mel-frequency cepstrum coefficients

Takahiro FUKUMORI†, Masato NAKAYAMA†, Takanobu NISHIURA†, and Hiroaki NANJO††

† College of Information Science and Engineering, Ritsumeikan University. 1-1-1 Nojihigashi, Kusatsu, Shiga, 525–8577, Japan.
†† Academic Center for Computing and Media Studies, Kyoto University. Nihonmatsu-cho, Yoshida, Sakyo-ku, Kyoto, 606–8501, Japan.
E-mail: †{fukumori@fc, mnaka@fc, nishiura@is}.ritsumei.ac.jp, ††nanjo@media.kyoto-u.ac.jp

Abstract　This paper describes a method based on combined features with mel-frequency cepstrum coefficients (MFCCs) and rahmonic in order to robustly detect a shouted speech in noisy-and-reverberant environments. MFCCs collectively make up mel-frequency cepstrum, and rahmonic shows a subharmonic of fundamental frequency in the cepstrum domain. In our previous method, Gaussian mixture models (GMM), hidden Markov model (HMM) and deep neural network (DNN) are constructed with the proposed features extracted from training data which includes a lot of normal and shouted speech samples. Especially, our latest study showed the effectiveness of our proposed method through detection experiments for shouted speech in clean and noisy environments. In this study, we evaluate the detection performance of shouted speech in noisy-and-reverberant environments. The results show that MFCCs and rahmonic were effective for representing an utterance mechanism including both vocal tract and vocal cords, and these features were independent of noise and reverberation. In addition, DNN could achieve higher performance in noisy-and-reverberant environments than GMM and HMM.

Key words　Shouted speech detection, Noisy-and-reverberant environment, Rahmonic, Mel-frequency cepstrum coefficients

1. はじめに

異常事態を検知する防犯システムは，人々に安全な生活環境を提供するために重要な役割を果たしている．異常事態には，視覚的な情報（不審な行動など）と聴覚的な情報（叫び声や異常音など）が含まれているが，現在の一般的な防犯システムは，カメラなどで撮影した動画像情報だけを利用して生活環境を監視している [1], [2]．この問題を解決するために，近年はカメラで計測した画像情報以外にマイクロホンで計測した音情報から異常事態を検出する研究が注目されている [3]～[5]．特にカメラの死角の状況を捉えられる音情報を現行の防犯システムに搭載することで，防犯システム周辺の音環境を理解することが可能となり，異常事態の検知性能を飛躍的に向上させられると期待できる．

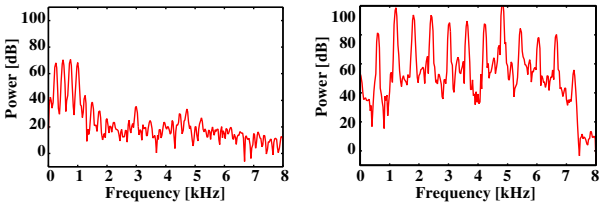
音情報を使って異常事態を検知する手法として，これまでに生活環境音の中から非日常的な音声である叫び声を検出する手法が数多く提案されてきた [6]～[9]．しかし，これらの手法には発話内容や評価環境の SNR に大きく依存するという問題があった．発話内容や雑音環境の SNR に依存しないアプローチとして，メルケプストラム（Mel-frequency cepstral coefficients: MFCC）に基づいて構築した GMM（Gaussian Mixture Model）を用いる手法 [10], [11] があり，雑音環境下で頑健に叫び声を検出できることが報告されている．メルケプストラムは音声の発声機構の中でも特に声道情報を重点的に表現しているが，ここで更に声帯情報に関わる音声特徴量を加味しながら叫び声を分析することで叫び声検出性能の向上が期待できる．

我々は，これまでに Rahmonic と呼ばれる基本周波数の低調波成分が叫び声検出に有効であることを明らかにし，従来のメルケプストラムと併用しながら叫び声を検出する方法を提案した [12]．具体的には，大量の平静音声と叫び声から抽出した Rahmonic とメルケプストラムに基づく音響モデル（Gaussian Mixture Model：GMM，Hidden Markov Model：HMM，Deep Neural Network：DNN）を用いて叫び声を検出していた．特に前報では，クリーン環境と雑音環境における叫び声の検出性能を評価し，提案手法（Rahmonic とメルケプストラムに基づいて構築した DNN）の高い検出性能を確認した．本稿では，更に実環境を想定して，雑音と残響が混在する環境において叫ばれた音声の検出性能を評価する．

2. Rahmonic とメルケプストラムを用いた叫び声検出

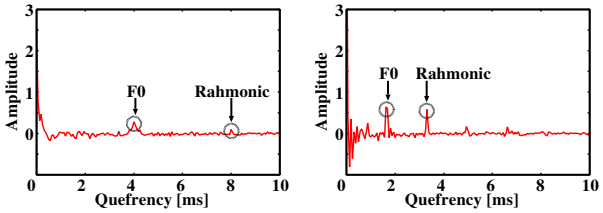
2.1 音声特徴量（メルケプストラム・Rahmonic）

我々は，これまでに Rahmonic とメルケプストラムを用いた叫び声検出法を提案した [12]．メルケプストラムは，人間の聴覚特性を考慮したケプストラム係数であり，音声認識では音韻を特定するための声道特徴量として用いられている [13]．一方，Rahmonic は，基本周波数の低調波成分であり，人間の声帯運動に関わる特徴を表現する [14]．そして，従来研究 [12] において，これらの音声特徴量が平静音声と叫び声で異なることが報告されている．



(a) 平静音声　(b) 叫び声

図 1　平静音声と叫び声の対数パワースペクトル



(a) 平静音声　(b) 叫び声

図 2　平静音声と叫び声のケプストラム

ここで，図 1 と図 2 に平静音声と叫び声に対する対数パワースペクトルとケプストラムを示す．まず対数パワースペクトルに着目すると，図 1(a) の平静音声よりも図 1(b) の叫び声の調波成分が強調されて表れていることが確認できる．またケプストラムにおいても，図 2(a) の平静音声には顕著に表れなかった Rahmonic を図 2(b) の叫び声では明確に確認することができる．このように周波数領域やケプストラム領域でも平静音声と叫び声の間に差異があることからも，メルケプストラムや Rahmonic を用いることで高精度に叫び声を検出することが期待できる．

2.2 検出アルゴリズム

図 3 に叫び声の検出手順を示す．叫び声を検出する方法として，はじめに予め収録した平静音声と叫び声から抽出した Rahmonic とメルケプストラムを用いて音響モデルを構築する．次に，実際の評価環境で収録した観測音声から Rahmonic とメルケプストラムを抽出し，これらの音声特徴量と学習した音響モデルを用いて観測音声を平静音声と叫び声のいずれかに分類する．

従来は音響モデルとして一般的に GMM が利用されていたが，我々はこれまでに叫び声検出に用いる音響モデルを従来の Gaussian Mixture Model (GMM) から Hidden Markov Model (HMM) や Deep Neural Network（DNN）に拡張して，それぞれの音響モデルが叫び声検出に与える影響を評価した [12]．GMM は観測音声に対する平均的な音声特徴量を用いて叫び声をモデル化している．HMM は音声特徴量の時間的変化を表現できる音響モデルであり，叫び声は平静音声と比べて発話時間やエネルギーの時間変動が異なること [15] から，HMM を用いることで音声特徴量の時間構造も考慮することで叫び声検出の性能改善が期待できる．そして，DNN はニューラルネットワークの 1 つであり，ネットワーク内で深い層構造を有しており，特に入力層を音響特徴量（本稿の場合，メルケプストラムや Rahmonic），出力層を発話様式（平静音声と叫び声）とし

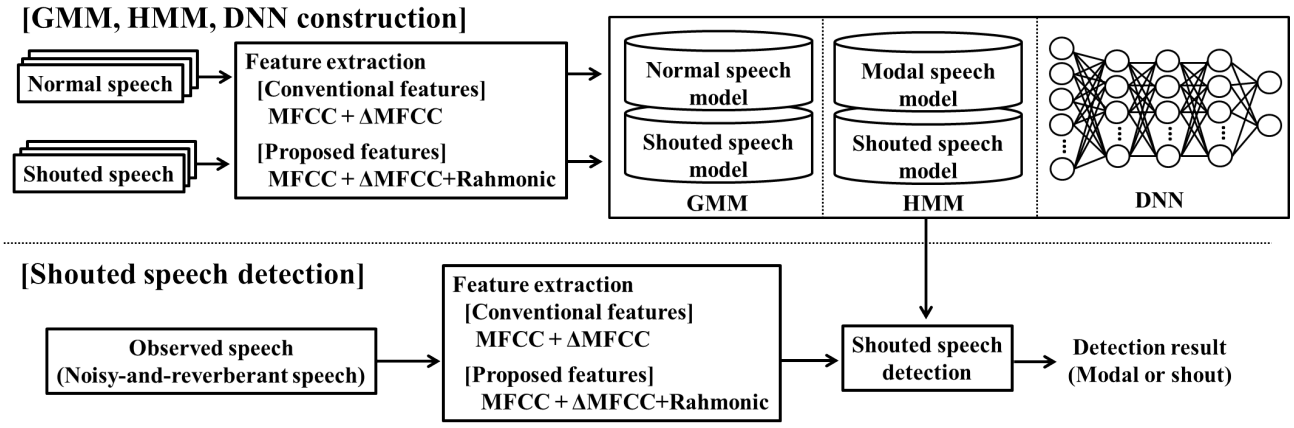


図 3　叫び声検出アルゴリズムの概要

て対応付けることで，DNN を叫び声検出のための音響モデルとして使用することができる．またネットワークに入力された音響特徴量に対して重み付けを行いながら出力層まで伝搬する過程は，評価環境に依存せずに叫び声検出に有効な特徴を重点的に抽出できると考えられる．

前報 [12] までは，クリーン環境と雑音環境における叫び声の検出実験を通して提案手法の有効性を確認していた．本稿では，更に実環境を想定して，雑音・残響下における叫び声検出性能を評価する．具体的には，雑音と残響が混入した音声を計算機上で生成し，その音声でマルチコンディション学習を行った音響モデルを用いて叫び声を検出する．

3. 評 価 実 験

3.1 実 験 条 件

本実験では，クリーン音声 (男女各 400 発話) に残響を畳み込んだ音声に雑音を加算した学習音声を用いて性別依存のマルチコンディション音響モデル (GMM，3 状態の HMM，DNN) を構築した．また前報における評価実験結果より，GMM と HMM の混合数は 128 を採用して評価を行った．また GMM と HMM の構築には HTK [16] を，DNN の構築には Chainer [17] を用いた．HMM は，left-to-right で 3 状態を遷移するモデルを構築した．DNN で用いる各音響特徴量の統合フレーム数は，1 フレーム（現在フレームのみ），7 フレーム（前後 3 フレームを含む），11 フレーム（前後 5 フレームを含む）の 3 種類，隠れ層は 3 層（各層の素子数は 20）とした．また発話様式の識別では，話者オープンテストを想定して音響モデルの学習で用いた音声とは異なる話者音声を用いた．雑音は，NOISEX-92 [18] よりホワイトノイズとスピーチバブルノイズとし，SNR は ∞, 20, 10, 0 dB の 4 種類を採用した．なお，各雑音から異なる 2 区間の信号を抽出し，それぞれを学習データとテストデータとした．残響は，3 種類の残響環境（和室：T_{60}=450 ms，会議室：T_{60}=600 ms，エレベータホール：T_{60}=850 ms）のインパルス応答を学習データに，3 種類の残響環境（研究室：T_{60}=450 ms，廊下：T_{60}=600 ms，階段：T_{60}=850 ms）のインパルス応答をテストデータとした．音声特徴量は，20 ms の分析フレームから MFCCs（12 次元），ΔMFCCs（12 次元），Rahmonic

表 1　実 験 条 件

Training data	Female speaker: 400 samples Male speaker: 400 samples
Testing data	Female speaker: 100 samples Male speaker: 100 samples
Sampling	16 kHz / 16 bit
Acoustic feature	12 orders MFCCs 12 orders ΔMFCCs 1 order Rahmonic
Acoustic model	1. GMM 2. HMM (3 states) 3. DNN
Noise	White noise, Speech babble [18]
Reverberation time (T_{60})	450 ms (Jpn. style room, Laboratory) 600 ms (Conference room, Corridor) 850 ms (Lift station, Stairs)
SNR	0, 10, 20, ∞ dB
Frame length	25 ms (Hamming window)
Frame shift	10 ms

（1 次元）を算出した．各手法の有効性を評価するための指標として，

1. 識別率 (%)
全ての平静音声と叫び声の内，正しく発話様式が識別された音声サンプル数の割合

2. 誤検出率 (%)
全ての平静音声の内，誤って叫び声として識別された音声サンプル数の割合

3. 誤棄却率 (%)
全ての叫び声の内，誤って平静音声として識別された音声サンプル数の割合

の 3 種類を用いた．また本実験で用意できた評価音声が少量であることを考慮して，今回は 5 分割交差検定を実施した．

3.2 実 験 結 果

図 4 に，全評価環境における音響モデルと音声特徴量ごとの平均識別率を示す．また表 2，3 にホワイトノイズにおける識別率，誤検出率，誤棄却率を，表 4，5 にスピーチバブルノイズにおける識別率，誤検出率，誤棄却率を示す．なお誤検出率と誤

表 2 Identification accuracy [%] in noisy-and-reverberant environments (Noise: Whitenoise).

GMM	SNR=∞ dB			SNR=20 dB			SNR=10 dB			SNR=0 dB		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	95.9	95.4	96.1	96.5	94.1	97.5	96.9	93.6	97.7	96.8	77.0	96.8
T_{60}=600 ms	93.9	93.6	95.4	97.0	94.1	97.4	97.5	93.7	97.3	97.5	76.3	97.5
T_{60}=850 ms	93.7	95.2	95.3	96.6	80.7	96.5	95.8	94.2	95.6	96.5	78.6	96.5

HMM	SNR=∞ dB			SNR=20 dB			SNR=10 dB			SNR=0 dB		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	92.2	87.6	95.0	96.3	91.1	98.0	96.3	92.0	97.1	92.6	84.1	97.3
T_{60}=600 ms	91.7	87.7	94.6	96.8	92.6	98.0	95.9	91.6	96.8	93.1	85.6	96.8
T_{60}=850 ms	91.2	75.1	94.8	96.4	82.4	97.0	94.9	76.5	96.2	93.2	77.8	96.4

DNN (1 frame)	SNR=∞ dB			SNR=20 dB			SNR=10 dB			SNR=0 dB		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	91.5	53.7	93.3	94.3	49.3	95.3	92.4	49.9	92.4	85.8	58.7	87.5
T_{60}=600 ms	92.4	53.5	93.4	94.5	52.6	96.2	94.6	50.3	95.8	90.6	63.1	90.5
T_{60}=850 ms	92.3	58.4	94.2	94.3	54.8	95.5	91.2	52.2	93.1	87.2	65.0	87.6

DNN (7 frames)	SNR=∞ dB			SNR=20 dB			SNR=10 dB			SNR=0 dB		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	96.6	60.6	97.8	99.5	55.9	99.4	98.8	60.5	98.3	95.6	46.0	97.6
T_{60}=600 ms	98.0	59.1	98.2	99.1	58.0	98.6	99.1	62.6	99.9	97.4	52.6	97.2
T_{60}=850 ms	98.4	66.2	99.4	99.0	65.3	99.6	99.1	59.8	99.8	93.8	52.1	95.4

DNN (11 frames)	SNR=∞ dB			SNR=20 dB			SNR=10 dB			SNR=0 dB		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	98.2	63.0	**99.1**	**99.8**	60.7	99.7	**99.2**	63.1	99.0	94.2	48.4	**95.6**
T_{60}=600 ms	**98.9**	59.8	**98.9**	**99.6**	62.9	99.3	**100**	62.5	**100**	97.6	45.6	**99.6**
T_{60}=850 ms	99.3	68.3	**99.6**	**99.9**	65.9	99.2	**99.8**	62.4	99.6	95.5	49.0	**97.3**

*M: MFCCs, R: Rahmonic, M+R: MFCCs and Rahmonic

表 3 Misdetection rate [%] and misrejection rate [%] in noisy-and-reverberant environments (Noise: Whitenoise).

Misdetection rate	**GMM**			**HMM**			**DNN** (1 frame)			**DNN** (7 frames)			**DNN** (11 frames)		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	5.6	12.4	4.6	9.3	16.8	4.4	8.8	58.4	7.5	2.1	53.8	2.0	1.5	46.0	**1.7**
T_{60}=600 ms	3.8	12.6	4.6	9.1	15.8	4.2	7.8	56.4	4.5	2.0	49.3	0.6	**0.3**	47.0	**0.3**
T_{60}=850 ms	6.1	31.4	7.4	9.2	37.2	5.6	12.5	56.0	8.0	3.5	42.0	0.6	3.3	36.5	**0.5**

Misrejection rate	**GMM**			**HMM**			**DNN** (1 frames)			**DNN** (7 frames)			**DNN** (11 frames)		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	1.0	11.2	1.1	**0.6**	5.1	0.7	9.5	43.9	8.5	2.0	44.5	1.3	2.3	41.8	2.0
T_{60}=600 ms	1.5	11.3	0.6	0.8	4.3	1.3	6.6	41.5	6.2	1.3	40.0	1.6	1.0	42.0	**0.4**
T_{60}=850 ms	1.3	10.8	**1.0**	1.1	5.0	1.2	8.7	39.9	7.2	2.6	42.7	2.0	1.4	41.8	1.5

*M: MFCCs, R: Rahmonic, M+R: MFCCs and Rahmonic

棄却率は，各雑音環境の SNR に対する平均結果を示している．表中の太字は各環境における最高性能を，そして「M」,「R」,「M+R」は，それぞれメルケプストラム，Rahmonic，両特徴量併用の結果を示す．

まず音響モデルに着目すると，表 2~4 より，DNN（特に統合フレーム数が 11 のとき）を音響モデルとして用いることで最も高い識別性能（雑音や残響の種類や SNR などに関係なく全ての環境で 95%以上の識別率）を達成した．これは DNN が GMM や HMM と比較して雑音や残響の影響を受けずに叫び声検出に有効な特徴を重点的に抽出できたためだと考えられる．また DNN の統合フレーム数については，前後 5 フレーム（計 11 フレーム）を用いることで，他のフレーム数を用いた場合よりも性能が改善したことから，叫び声検出には時間的構造も考慮した入力特徴量が有効であることがわかった．

次に入力特徴量に着目すると，全体的にメルケプストラム単体，あるいはメルケプストラムと Rahmonic を併用した特徴量を用いることで高い識別性能を達成することができた．特に SNR が低い環境（SNR= 0 dB）ほど，メルケプストラムと Rahmonic を併用した方が識別性能が改善することを確認した．この結果からも雑音や残響が混在する環境においても，人間の発声機構（声道特性と声帯特性）をそれぞれメルケプストラムと Rahmonic を用いて効率よく表現できていると考えられる．

表 4 Identification accuracy [%] in noisy-and-reverberant environments (Noise: Speech babble).

GMM	SNR=∞ dB			SNR=20 dB			SNR=10 dB			SNR=0 dB		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	97.9	92.4	97.1	97.6	92.1	98.7	94.6	91.6	97.3	88.7	75.1	91.6
T_{60}=600 ms	93.9	90.6	91.4	97.7	91.0	96.7	95.9	91.2	97.0	88.3	74.4	91.5
T_{60}=850 ms	93.7	75.2	93.3	97.2	79.5	97.6	95.2	83.3	95.2	90.3	68.5	89.2

HMM	SNR=∞ dB			SNR=20 dB			SNR=10 dB			SNR=0 dB		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	94.2	77.6	94.0	97.5	77.0	96.5	94.4	74.8	94.2	91.2	74.9	83.6
T_{60}=600 ms	93.7	77.7	94.6	97.8	77.1	96.2	95.0	75.2	93.6	95.2	74.2	82.3
T_{60}=850 ms	93.2	75.1	94.8	97.3	78.0	96.9	95.3	74.5	90.9	94.4	68.6	82.0

DNN (1 frame)	SNR=∞ dB			SNR=20 dB			SNR=10 dB			SNR=0 dB		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	91.5	53.7	93.3	89.7	51.9	91.2	85.5	54.3	88.8	79.5	66.6	79.2
T_{60}=600 ms	92.4	53.5	93.4	89.3	54.0	90.2	85.0	54.8	86.3	80.7	67.0	81.4
T_{60}=850 ms	92.3	58.4	94.2	87.7	58.2	88.9	83.7	61.0	86.0	79.0	73.8	81.9

DNN (7 frames)	SNR=∞ dB			SNR=20 dB			SNR=10 dB			SNR=0 dB		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	96.6	60.6	97.8	96.1	58.1	96.6	94.7	62.8	96.6	91.4	69.7	91.3
T_{60}=600 ms	98.0	59.1	98.2	96.0	61.4	96.4	94.0	62.0	95.4	90.4	69.8	94.3
T_{60}=850 ms	98.4	66.2	99.4	96.1	67.0	96.7	94.1	70.0	95.7	88.0	63.7	91.3

DNN (11 frames)	SNR=∞ dB			SNR=20 dB			SNR=10 dB			SNR=0 dB		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	98.2	63.0	**99.1**	97.6	60.9	**97.8**	95.2	64.7	**96.7**	92.8	71.1	**93.2**
T_{60}=600 ms	**98.9**	59.8	**98.9**	**96.6**	64.4	**96.6**	**96.1**	66.1	95.6	93.4	71.0	**93.9**
T_{60}=850 ms	99.3	68.3	**99.6**	**97.5**	67.7	96.8	96.1	72.7	**96.3**	92.6	70.9	**95.4**

*M: MFCCs, R: Rahmonic, M+R: MFCCs and Rahmonic

表 5 Misdetection rate [%] and misrejection rate [%] in noisy-and-reverberant environments (Noise: Speech babble).

Misdetection rate	**GMM**			**HMM**			**DNN** (1 frame)			**DNN** (7 frames)			**DNN** (11 frames)		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	11.8	11.8	4.7	7.2	38.1	16.1	15.1	60.0	11.6	4.3	51.2	2.3	2.7	49.5	**2.0**
T_{60}=600 ms	10.9	12.9	6.3	4.2	37.4	16.6	15.1	59.4	11.6	5.9	49.7	**1.8**	2.0	48.5	2.0
T_{60}=850 ms	10.6	33.2	8.0	4.7	49.0	19.6	18.9	54.6	12.1	12.1	42.2	**2.2**	6.5	35.9	3.2

Misrejection rate	**GMM**			**HMM**			**DNN** (1 frame)			**DNN** (7 frames)			**DNN** (11 frames)		
	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
T_{60}=450 ms	1.0	15.8	3.5	4.1	10.8	**1.0**	15.3	36.8	14.7	6.8	32.0	6.1	5.4	31.2	4.7
T_{60}=600 ms	1.1	16.0	3.5	3.7	11.5	**2.0**	15.3	35.9	15.0	6.1	31.2	6.7	5.3	29.8	5.4
T_{60}=850 ms	1.0	12.6	4.0	3.0	3.6	**1.4**	16.2	31.0	15.0	7.0	31.1	5.2	4.9	29.1	4.2

*M: MFCCs, R: Rahmonic, M+R: MFCCs and Rahmonic

最後に誤検出率と誤棄却率に注目すると，ホワイトノイズよりもスピーチバブルノイズの方が，全体的に誤検出率と誤棄却率が高くなった．これはスピーチバブルノイズが音声に近い特徴を有していることが，平静音声と叫び声を識別する際に影響を与えたものだと考えられる．またいずれの環境においても，全体的に誤検出率が誤棄却率を上回る結果であったことから，雑音や残響が発話音声に混入することで平静音声が叫び声に誤識別されやすいことがわかった．

以上のことより，メルケプストラムと Rahmonic に基づいて構築した DNN が叫び声検出に有効であることを確認できた．今後は本提案手法と雑音・残響抑圧手法を組み合わせて，入力音声から雑音や残響の影響を取り除くことで更なる識別率の改善が期待できる．

4. おわりに

本稿では，雑音・残響環境下における Rahmonic とメルケプストラム（Mel-Frequency Cepstrum Coefficients: MFCCs）を用いた叫び声検出を行った．実験結果より，雑音や残響の種類や SNR に依らず，Rahmonic とメルケプストラムを併用することで高い叫び声検出性能を実現することができた．また叫び声検出に有効な特徴を重点的に抽出できる DNN を用いることで 90 %以上の叫び声検出性能を達成した上に，従来の GMM

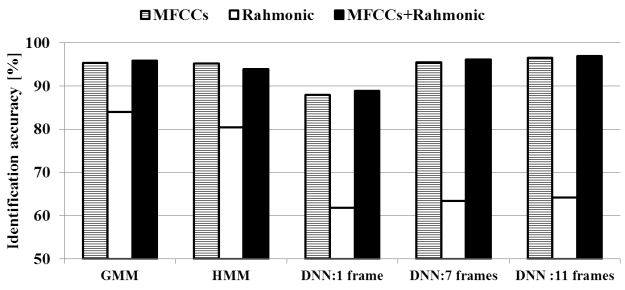


図 4　音響モデルごとの平均識別率

や HMM よりも叫び声の検出性能が改善した．今後は，本提案手法と雑音・残響抑圧手法を組み合わせて，より外乱に頑健な叫び声検出に取り組む計画である．

謝辞　本研究の一部は，科研費（16K16094）の研究助成を受けた．

文　献

[1] W. Yao-Dong, T. Takeshi, and I. Idaku, “HFR-video-based machinery surveillance for high-speed periodic operations,” *Journal of System Design and Dynamics*, vol. 5, no. 6, pp. 1310-1325, 2011.

[2] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, “Scream detection for home applications,” *5th IEEE Conference on Industrial Electronics and Applications*, pp. 2115-2120, 2010.

[3] M. Cowling, “Comparison of techniques for environmental sound recognition,” *Pattern Recognition Letter*, vol. 24, no. 15, pp. 2895-2907, 2003.

[4] K. M. Kim, J. W. Jung, S. Y. Chun, and K. S. Park, “Acoustic intruder detection system for home security,” *IEEE Transaction on Consumer Electronics*, vol. 51, no. 1, pp. 130-138, 2005.

[5] K. Hayashida, J. Ogawa, M. Nakayama, T. Nishiura, and Y. Yamashita, “Multi-stage identification for abnormal/warning sounds detection based on maximum likelihood classification,” *ICA2013*, PaperID:1pSPb4, 2013.

[6] J. L. Rouas, J. Louradour, and S. Ambellouis, “ ,” *IEEE Intelligent Transportation Systems Conference*, pp. 733-738, 2006.

[7] P. K. Atrey, N. C. Maddage, and M. S. Kankanhalli, “Audio based event detection for multimedia surveillance,” *ICASSP 2006*, pp. 813-816, 2006.

[8] S. Ntalampiras, I. Potamitis, and N. Fakotakis, “An adaptive framework for acoustic monitoring of potential hazards,” *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[9] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” *IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 21-26, 2007.

[10] J. Pohjalainen, P. Alku, and T. Kinnunen, “Shout detection in noise,” *ICASSP 2011*, pp. 4968-4971, 2011.

[11] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, “Scream detection for home applications,” *Industrial Electronics and Applications 2010*, pp. 2115-2120, 2010.

[12] 福森 隆寛, 中山 雅人, 西浦 敬信, 南條 浩輝, “Rahmonic とメルケプストラムを用いた深層ニューラルネットワークによる叫び声検出の検討,” 日本音響学会 2017 年春季研究発表会, pp. 125-126, 2017.

[13] J. Benesty, M. M. Sondhi, and Y. Huang, “Springer handbook of speech processing,” *Springer*, 2008.

[14] A. M. Noll, “Cepstrum Pitch Determination,” *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 203-309, 1967.

[15] C Zhang and J.H.L Hansen “Analysis and classification of speech mode: whispered through shouted,” *INTERSPEECH 2007*, pp. 2289-2292, 2007.

[16] HTK Software Toolkit, http://htk.eng.cam.ac.uk/

[17] Chainer, https://chainer.org/

[18] A. Varga and H.J.M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247-251.